Ügyfél érzelmi állapotának detektálása telefonos ügyfélszolgálati dialógusban

Vicsi Klára, Sztahó Dávid
Budapesti Műszaki és Gazdaságtudományi Egyetem

Távközlési és Médiainformatikai Tanszék Beszédakusztiai Laboratórium,
1111 Budapest, Sztoczek utca 2.
vicsi@tmit.bme.hu, sztaho@tmit.bme.hu

Kivonat: A cikkünkben egy érzelem-felismerési kísérletről számolunk be, ahol a spontán társalgás során a semlegesről idegesre, feszültre megváltozott érzelmi állapotot kívánjuk automatikusan detektálni, telefonon keresztül. A cél egy automatikus figyelőrendszer kifejlesztése, amely meghatározza az ügyfél elégedettségének, vagy elégedetlenségének a mértékét. Ehhez a munkához létrehoztuk, 1000 telefonhívás-felvételből az ún Magyar Telefonos Ügyfélszolgálati Beszéd Adatbázist (MTÜBA), amelyben a spontán dialógusok nyelvi tartalmát, valamint frázisonkénti érzelmi tartamát jelöltük be. Az akusztikai előfeldolgozás után az érzelem-felismerést support vector machine (SVM) osztályozó segítségével végeztük. Az SVM osztályozóval végül is csak 2 állapotot, egy semleges, és egy elégedetlenséget kifejező (ideges és panaszkodó együtt) állapotot különböztettünk meg. Az automatikus figyelőrendszer részére kiválasztottunk 15 másodperc hosszú figyelő ablakot, amelyen belül összeszámoltuk az elégedetlenséget jelző frázisok számát. Ez adta meg az elégedetlenség mértékét. Az ablakot 10 másodpercenként léptettük előre a beszélgetés folyamán. Kísérletezéssel beállítható volt egy olyan elégedetlenségi mérték küszöb, amely felett jelzés (riasztás) történik. Amennyiben ez a küszöb a 30%-os elégedetlenségi mérték, akkor az átlagos riasztási pontosság 89,6% volt, ami legtöbbször csak a kézi és az automatikus riasztás közötti időcsúszásból eredt. Így a kifejlesztett automatikus figyelőrendszer hasznos eszköz lehet diszpécser központokban.

1 Bevezetés

Az emberi beszédkommunikációban a beszéd információfeldolgozásának két egymástól elkülönült feldolgozási módjáról beszélhetünk. Az egyik feldolgozási mód esetében speciális szemantikai tartalmú üzeneteket dolgozunk fel (verbális csatorna); a másik információfeldolgozási mód az, ahol a beszélő általános érzelmi, egészségi állapotát, hangulatát dolgozzuk fel (a nem verbális csatorna) [1]. Az utóbbi évtizedekben óriási erőfeszítések történtek a verbális csatorna működésének megértésére. A nem verbális csatorna jelentősége ez ideig kisebb volt, és működését kevésbé értjük.

Az emberi beszéddel nagyon sok mindent ki lehet fejezni a nyelvi tartalmon kívül, amelyeket különböző beszédváltozatok jelenítenek meg, például a beszédstílus, rit-

mus, hangerő, hangszín, intonáció – ezek mind széles körben használatosak arra, hogy a beszélő érzelmi, egészségi állapotát egyidejűleg kifejezzék. Csak az utóbbi években növekedett meg a jelentősége a beszéd különböző paralingvisztikai és extralingvisztikai nézőpont szerinti vizsgálatának. Az irodalomban található néhány kutatási leírás, amely a beszéd érzelemtartalmának vizsgálatával, és az érzelem automatikus felismerésével foglalkozik, de ezek az eredmények mind laboratóriumi körülmények között elhangzó tiszta beszédre vonatkoznak [2, 3, 4, 5]. A publikációk legtöbbjében szimulált különböző érzelemtartalmú beszédet használnak, leggyakrabban művészek bemondásmintáit. Az érzelem jellemzésére a pszichológiában, nyelvészetben és audiovizuális jelfeldolgozásban, például az MPEG-4 szabvány leírásában [6] hagyományos érzelemkategóriákat használnak, úgymint boldogság, szomorúság, düh, meglepetés, undor. Eredetileg az MPEG-4 szabványban e kategóriákat az arcmimika jellemzésére szolgáló virtuális paraméterek (facial animation parameters, FAPs) megjelenítésére használták.

A valóságban rendszerint spontán beszédet használunk, és a spontán beszédre jellemző adatok igen nagymértékben különböznek a színészek által produkált szép beszédtől [7], és a beszédtechnológiai alkalmazásokban a valóságos spontán beszéd alkalmazása az, ami szükséges. Az utóbbi években már megjelent néhány olyan publikáció, amely a spontán hétköznapi beszéd vizsgálatával [8] és információtartalmainak felismerésével [9] foglalkozik.

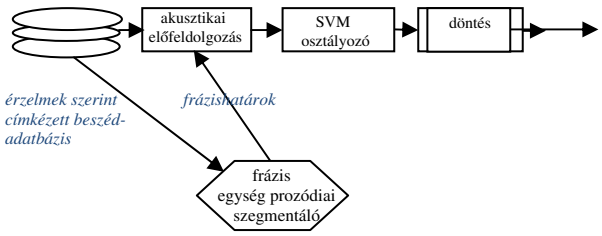
Jelen cikkünkben a telefondiszpécser és az ügyfél közötti hétköznapi spontán társalgási adatbázis alapján végzett automatikus érzelem-felismerési kísérletekkel foglalkozunk. Az akusztikai előfeldolgozásnál támaszkodtunk a korábban végzett, imitált érzelemtöltetű beszéd felismerési kísérleteink eredményeire [10].

A cikkünkben a beszéd érzelmet kifejező akusztikai paramétereinek a felismerését tárgyaljuk, de tervezzük a verbális csatornán keresztül is a nyelvi tartalom érzelemre vonatkozó statisztikai jellegzetességeinek vizsgálatát is.

2 Rendszerleírás

Egy beszélgetés során, különösen, ha az hosszan tartó, a beszélő érzelmi állapota, hangulata változik. Ha követni akarjuk a beszélő érzelmi változásait, szegmensekre kell felosztanunk a beszédfolyamot, így meg tudjuk vizsgálni, hogyan változik szegmensről szegmensre a beszélő érzelmi állapota a beszélgetés alatt. Rendszerünkben a frázist választottuk szegmentálási egységként, a korábbi tanulmányaink során nyert tapasztalatok alapján [10]. A frázis méretű egységek szegmentálásakor az egységekre való osztást prozódiai szegmentálónk végezte el [11]. (Ezt a szegmentálót a folyamatos beszéd felismerés részeként a beszéd szemantikai feldolgozására fejlesztettük ki, amelyet a frázis- és mondathatárok detektálására és a modalitás (mondattípus) felismerésére használtunk.)

Az akusztikai előfeldolgozás után a frázis méretű szegmenseket azok érzelmi töltete szerint osztályoztuk, SVM (support vector machine) gépi osztályozót használva. A rendszerünk folyamatábráját az 1. ábra szemlélteti.



1. ábra. Beszédérzelem osztályozónk blokkvázlata.

Kezdetben négy különböző érzelmi állapot került megkülönböztetésre a rögzített dialógusokban: semleges (N), ideges (I), panaszkodó (P), és egyéb (E). Később, a kísérletek tapasztalata alapján ezeket az érzelmeket összevontuk, már csak összesen két érzelmi osztályt különböztetve meg.

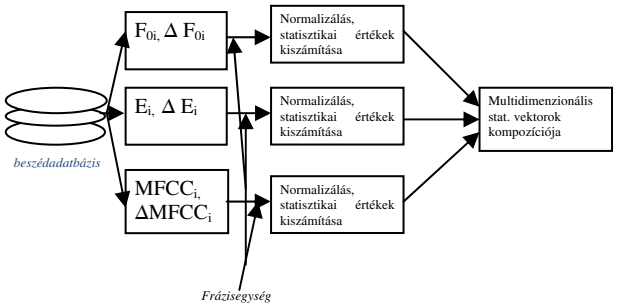
Végezetül ahhoz, hogy az érzelmi döntéshozás biztosabb legyen, egyszerre több frázis együttes kezeléséből alkotunk végleges döntést a beszélő érzelmi állapotáról.

2.1 Akusztikai előfeldolgozás

Általánosságban az alapfrekvencia, az intenzitás és annak időbeli függése a leghagyományosabban használt fizikai jellemző az érzelmek kifejezésére, mind a beszédfelismerés, mind a beszédszintézis területén. Azonban a korábbi automatikus beszéd felismerési kísérleteink során kiderült, hogy spektrális információ hozzáadása nagymértékben javítja az érzelem-felismerési eredményeket [10]. Ennek megfelelően az alapfrekvenciákat (F_{0i}), az intenzitásértékeket (E_i), 12 MFCC-t és deriváltjaikat mértük, 150 ms időablakot használva 10 ms időkeretekben, összesen 28 tulajdonságvektorral 10 ms-ként. Ezután a frázis prozódiai szegmentáló kijelöli a frázishatárokat a beszédben, frázisok sorozatát hozva ezzel létre. A 10 mszekundumonkénti tulajdonságvektorok alapján minden egyes frázist egy multi-dimenzionális statisztikai tulajdonságvektor jellemez, amint azt a 2. ábra mutatja. Ezeket a statisztikai tulajdonságvektorokat a következők szerint számítottuk ki: először F_{0i} értekeit az első időkeret F_{0i} értékeire, az E értékeket pedig az E maximum érték szerint normalizáltuk minden egyes frázis esetében. Majd e normalizált paraméterekből számítottuk ki a következő statisztikai adatokat minden egyes frázisnál:

- F_{0i} maximum, minimum, közép, medián értékei
- ΔF_{0i} maximum, minimum, közép, torzulás (skew) értékei
- E_i közép, medián értékei
- ΔE_i maximum, minimum, közép, torzulás (skew) értéke

- $MFCC_i$ maximum, minimum, közép értékei
- $\Delta\ MFCC_i$ maximum, minimum, közép értékei



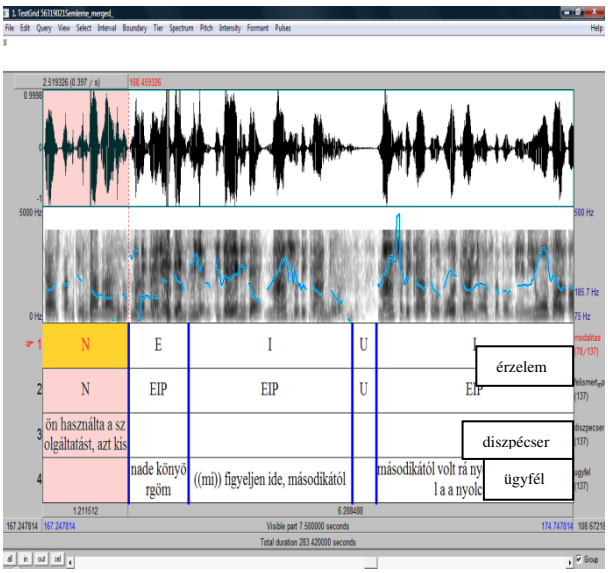
2. ábra. Akusztikai előfeldolgozás

2.2 Telefonos Ügyfélszolgálati Beszéd Adatbázis (TÜBA)

A TÜBA egy telefonos ügyfélszolgálat dialógusainak gyűjteménye, amely telefonvonalon keresztül lett rögzítve, 250-3500 Hz közötti frekvenciasávban, 8000 Hz-es mintavételi sebességgel és 16 bites amplitúdó felbontásban. A diszpécserek és ügyfelek közötti párbeszédek időtartalma változó, 1 és 30 perc közötti volt. A hanganyag feldolgozásához, a szegmentáláshoz és a címkézéshez a közismert Praat fonetikai feldolgozó programot [13] használtuk, mivel ez az eszköz megfelelő a párhuzamos feldolgozáshoz. A frázishatárok bejelölése után frázisonként bejegyzésre került a nyelvi tartalom, és a hozzá tartozó érzelem is párhuzamosan. A beszélő, a beszélő neme szintén bejegyzésre került.

A frázishatárok automatikus kijelölésére a prozódiai szegmentálónkat [11] használtuk , amint azt már az előző fejezetben is említettük. Azután szakértők kézzel javították a határokat, érzelem szerint felcímkézték a frázisszegmenseket. Négy különböző érzelmi állapotot különböztettek meg a rögzített párbeszédekben: semleges (N), ideges (I), panaszkodó (P), és egyéb (E). Gyakorlatilag nem volt több érzelemtípus az 1000 hívásban, csupán ez a négy. Sajnos sok esetben az ügyfél beszéde semleges volt. Összesen 346 ideges, 603 panaszkodó, és 225 egyéb frázis volt az ügyfelek beszédében, valamint több ezer semleges, amelyből 603 tipikusan semleges frázist választottunk ki a négy érzelem betanítására az osztályozási kísérletben.

A párbeszédek szegmentálásának és címkézésének egy példáját a 3. ábra mutatja be. A kézi szegmentálás és címkézés a harmadik sorban jelenik meg, osztályozónk címkézési eredménye pedig alatta látható. Az ügyfél és a diszpécser beszédének szövege a beszéddel és az érzelemmel párhuzamosan került lejegyzésre.



3. ábra. Példa a TÜBA szegmentálására és címkézésére. U: szünet, N: semleges, I: ideges, P: panaszkodó és E: egyéb.

2.3 A rendszer tesztelése

Frázisok érzelem szerinti osztályozása

Érzelmi osztályozónk betanítására és tesztelésére az úgynevezett „leave-one-out cross-validation” (LOOCV) módszert használtuk [12], amely egyetlen frázist használ értékelési adatként, a hívás fennmaradó frázisait pedig betanítási adatként. Majd ez úgy ismétlődik, hogy végül is minden egyes frázis egyszer értékelési adatként kerül felhasználásra. Az 1. táblázat mutatja a négy érzelem esetében kapott hibamátrixot.

1. táblázat: E, I, P, N érzelmek felismerési hibamátrixa.

	E	I	N	P	Pontosság
E	**49**	26	62	88	22%
I	9	**153**	60	124	44%
N	14	38	**398**	153	66%
P	11	70	157	**365**	60%
				átlag	54%

Az I és P érzelmeket nemcsak az osztályozó, de az emberek is alig tudták differenciálni. Így az I, P és E osztályok egy osztályba kerültek, mint elégedetlenséget kifejező érzelmek. Tehát végül az „elégedetlen” osztályt és a semleges érzelmek osztályát

különböztettük meg, és így tanítottuk be az SVM osztályozót. A teszteredményeket a 2. táblázat szemlélteti.

2. táblázat: Az (E, I, P), mint elégedetlen érzelmű összevont osztály, és az (N) semleges érzelmű osztály felismerési hibamátrixa.

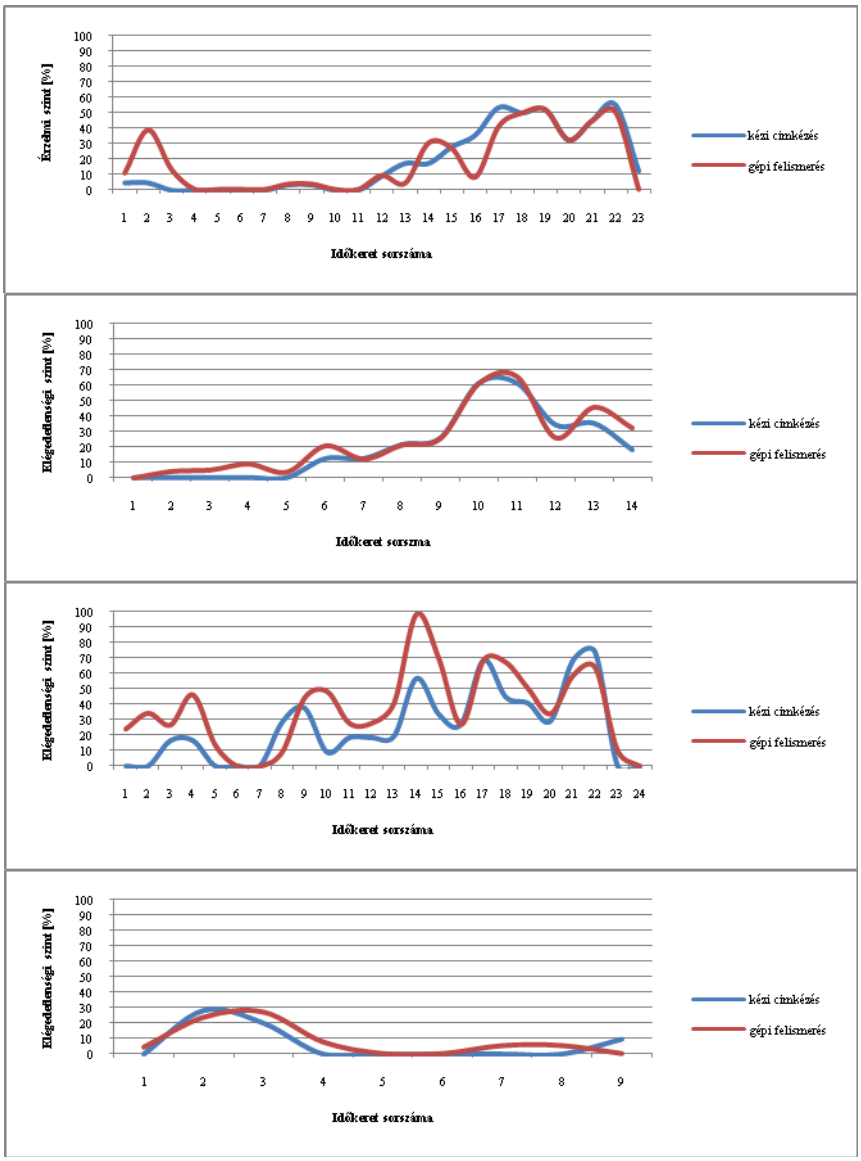
	EIP	N	Pontosság
EIP	**887**	287	76%
N	335	**839**	71%
		átlag	**73%**

Az ügyfél érzelmi állapotának detektálása

E kutatómunkának a célja annak a feltárása, hogy egy beszélgetés során hogyan lehetséges az ügyfelek érzelmi állapotát automatikusan felismerni. Frázisonként változhat, ugrálhat a megítélt érzelem. Biztos döntés akkor hozható, ha több frázison keresztül többségében egy típusú érzelem fordul elő. Ehhez előzetes kísérletezgetés alapján 15 másodperc hosszúságú időablakot választottunk, és mértük az ablakon belül az „elégedetlen"-nek osztályozott frázisok számát. Ez a szám %-ban kifejezve adta meg az „elégedetlenség" mértékét. (Az elégedetlenség akkor volt 100%-os, amikor a monitorozó ablakban az összes frázis elégedetlennek lett minősítve.) Azután az ablakot továbbmozgattuk, 10 másodperc időlépéssel. A 4. ábrán néhány példa jelenik meg arról, hogyan változik meg az ablakban mért szám, vagyis az elégedetlenség mértéke a beszélgetés során. Az automatikusan nyert eredményeket összehasonlítottuk a kézzel felcímkézett eredményekkel.

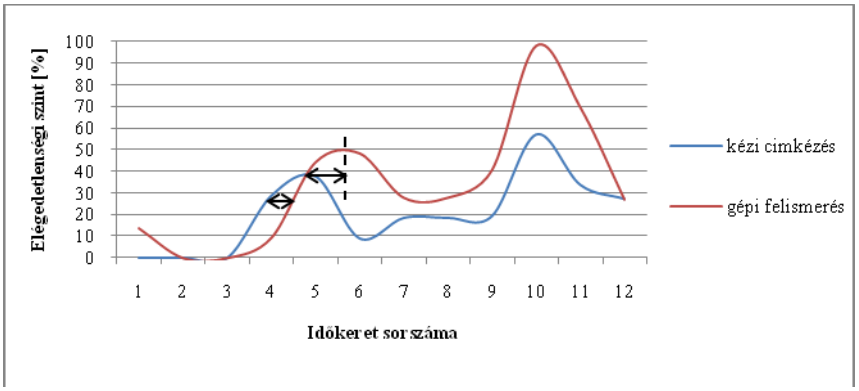
Egészében véve folyamatos megfigyelés esetében az automatikusan nyert, és a kézzel címkézett eredmények között az átlagos távolság 11,3% volt, összehasonlítva minden 10 másodperces időlépésben a megfigyelt eredményeket, és átlagolva a kapott különbségeket az egész adatbázishoz.

A valós felhasználásban az automatikus felismerés fő célja jelezni, amikor az elégedetlenségi szint elérte a kritikus szintet. Mi ezt „riadószint"-nek nevezzük. Ez a „riadószint" manuálisan beállítható. Például, válasszuk 30 százalékra a „riadószint"-et (ez azt jelenti, hogy 30% felett van a mért elégedetlenség). Vizsgáljuk meg ebben az esetben a riasztási pontosságot. Ezt úgy végezhetjük el, hogy összehasonlítjuk, az automatikus riasztást azzal a riasztással, amit az előzetesen kézzel címkézett anyagon számolunk. Az összehasonlítás keretről keretre történt. A különbségeket riadódetektálási hibának tekintettük. Az átlagos riadódetektálási hiba 10,4%-os volt.



4. ábra. Az ügyfél elégedetlenségének mértéke egy beszélgetés során. (Az elégedetlenség akkor volt 100%-os, amikor a monitorozó ablakban az összes frázis elégedetlennek lett minősítve.) Az automatikusan nyert eredményeket összehasonlítottuk a kézzel felcímkézett eredményekkel.

Ha csak azokat a párbeszédeket nézzük, ahol egyáltalán nem volt „riadószint” (semleges párbeszédek), a „riadószint” detektálási hiba 6,8%-os volt. Ez azt jelenti, hogy ha csak a több mint 30 százalékos elégedetlenséget tekintjük „elégedetlen” érzelmi állapotnak, az automatikus felismerési arány 93,2%-os. Egyéb párbeszédek érzelmi töltete (ahol a kézi felcímkézés legalább egy esetben elérte a „riadószint”-et) 77,2%-ban került felismerésre. A hibák fő oka az automatikus adatfelismerés és a kézi felcímkézés közötti kismértékű eltolódás. Ezt illusztrálja az 5. ábra.



5. ábra. A 4. ábra 3. diagramjának kinagyítása, példa a kézi felcímkézés és az automatikus felismerés görbéi közötti kismértékű eltolódásra.

3 Összegzés

A kísérletsorozat kezdetén, a 2.3.1. bekezdésben négy különböző érzelmi állapotot különböztettünk meg a rögzített párbeszédekben: semleges (N), ideges (I), panaszkodó (P), és egyéb (E). Ezeknek az érzelmeknek az átlagos osztályozási pontossága csupán 54%-os volt. Az osztályozási pontosság természetesen bizonyos mértékig növelhető a betanítási anyag növelésével, de az érzékelési kísérletek során, még művészek által előadott beszédnél is az emberi érzelem-felismerés (nem verbális csatornákon) általánosságban kevesebb volt, mint 70% (hat alapérzelem esetében) [2, 3, 5, 10] specifikus szemantikai tartalom nélkül (verbális csatornák). Ebből következik, hogy aligha várható sokkal jobb eredmény az automatikus érzelem-felismerés esetében, spontán beszédnél. Világos, hogy sokkal jobb eredmény érhető el, ha a verbális csatorna néhány információja a rendszerhez integrálódik. Ez az oka annak, hogy a lingvisztikai tartalmat is rögzítettük az adatbázis feldolgozáson keresztül, amint azt a 2.2 bekezdésben leírtuk. A jövőben azt tervezzük, hogy néhány lingvisztikai információt is feldolgozunk, és a két csatorna információit fogjuk integrálni.

A 2.3.2. bekezdésben leírt második kísérletünk során az osztályozott frázisokat egy időablakon keresztül figyeltük meg, hosszabb ideig, mint ameddig a frázis tart, hogy specifikusabb döntést hozhassunk a beszélő érzelmi állapotát illetően. Ez a megfigyelési technika képesnek látszik arra, hogy riasztást adjon, ha az ügyfél elégedetlensége

túlmegy egy bizonyos küszöbön, még verbális csatorna használata nélkül is. Ennek megfelelően a leírt döntési technika hasznos lehet a diszpécserközpontokban.

Köszönetnyilvánítás

Ezúton kívánunk köszönetet mondani az SPSS Hungary Ltd.-nek és az INVITEL Telecom Zrt.-nek a rendelkezésünkre bocsátott 1000 dialógusért.

Hivatkozások

1. Burkhardt, F., Paeschke A. et al.: A database of German Emotional Speech. N: Proc. Of Interspeech2005 (2005) 1517-1520
2. Campbell, N.: Getting to the heart of the matter. Keynote Speech in Proc. Language resources and Evaluation Conference (LREC-04), Lisabon, Portugal (2004)
3. Campbell, N.: Individual Traits of Speaking Style and Speech Rhythm in a Spoken Discourse. COST Action 2102 International Conference on Verbal and Nonverbal Features….Patras, Greece, (2007) 107-120
4. Douglas-Cowie, E. – Campbell, N. – Cowie, R. – Roach, P.: Emotional speech: towards a new generation of databases. Speech Communication 40. (2003) 33–60
5. Hozjan, V. – Kacic, Z.: A rule-based emotion-dependent feature extraction method for emotion analysis from speech. The Journal of the Acoustical Society of America. May, Vol. 119, Issue 5. (2006) 3109-31206
6. Kohavi, R.: "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12) (1995) 1137–1143
7. Kostoulas, T., Ganchev, T., Fakotakis, N.: Study on Speaker-Independent Emotion Recognition from Speech on Real-World Data, COST Action 2102 International Conference on Verbal and Nonverbal Features….Patras, Greece, October 2007. (2007) 235-242.8
8. Navas, E. – Hernáez, I. – Luengo, I.: An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS. IEEE Transaction on Audio, Speech, and Language Processing, vol. 14, no. 4, July, 2006 (2006)
9. MPEG-4: ISO/IEC 14496 standard. http://www.iec.ch, (1999)
10. Tóth Sz. L., Sztahó D., Vicsi K.: Speech Emotion Perception by Human and Machine. Proceeding of COST Action 2102 International Conference, Patras, Greece, October 29-31, 2007: Revised Papers in Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction 2008. ISBN: 978-3-540-70871-1. Springer LNCS (2008) 213-224
11. Praat, http://www.fon.hum.uva.nl/praat/
12. Vicsi, K. Szaszák, Gy.: Using Prosody for the Improvement of ASR: Sentence Modality Recognition. In: Interspeech 2008. Brisbane, Ausztrália 2008.09.23-2008.09.26. ISCA Archive, http:www.isca-speech.org/archive, (2008)
13. Wilting, J., Kramber, E., Swerts, M.: Realvs. Acted emotional speech.In:Proc. Of the Interspeech 2006 (2006) 805-808